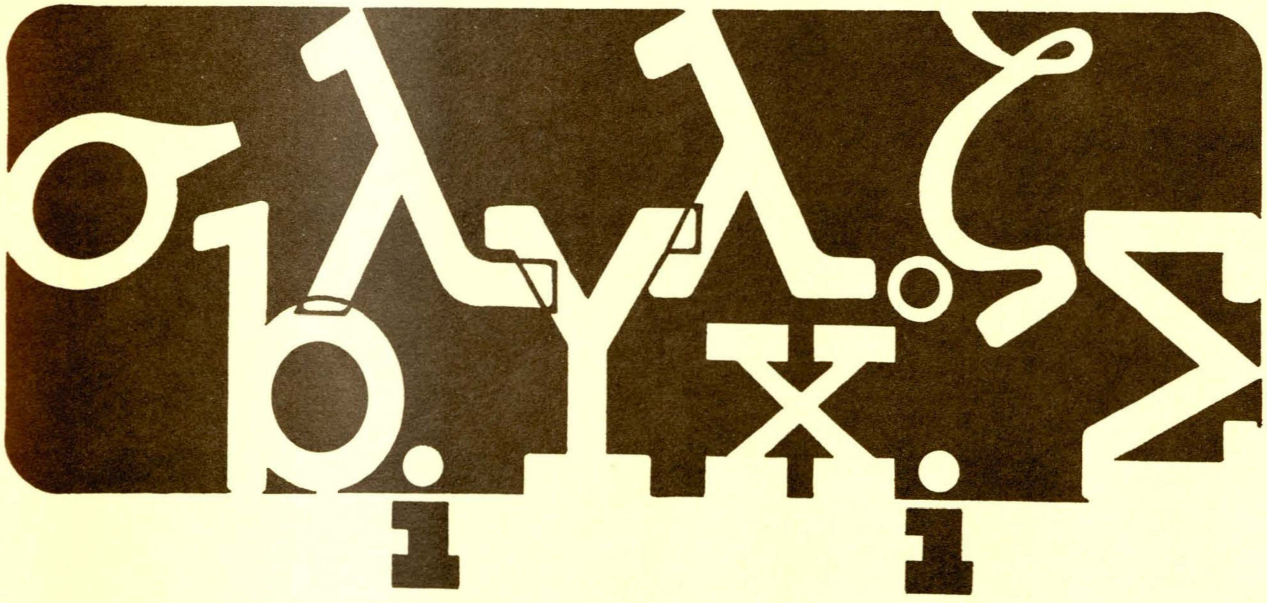
USDA FOREST SERVICE
GENERAL TECHNICAL REPORT NC-17

introduction to uses and interpretation of principal component analysis in forest biology

by J. G. Isebrands and Thomas R. Crow



NORTH CENTRAL FOREST EXPERIMENT STATION
FOREST SERVICE
U.S. DEPARTMENT OF AGRICULTURE

CONTENTS

Introduction 1
Principal Component Analysis 1
  Historical Development 1
  Basic Properties 2
  Terminology and Notation 3
  Operational Sequence 4
Examples of Applications 5
  Discarding Variables 5
  Ordinating Groups of Variables 8
  Multiplotting 11
  Principal Components in Conjunction
    With Regression Analysis 12
Literature Cited 17
Other References 19

THE AUTHORS: J. G. Isebrands is a Wood Anatomist for the Station at its Institute of Forest Genetics, Rhinelander, Wisconsin. Thomas R. Crow, formerly a Plant Ecologist at the Institute, is now located at the Institute of Tropical Forestry, Rio Piedras, Puerto Rico.

North Central Forest Experiment Station
John H. Ohman, Director
Forest Service - U.S. Department of Agriculture
Folwell Avenue
St. Paul, Minnesota 55101
1975

INTRODUCTION TO USES AND INTERPRETATION OF PRINCIPAL COMPONENT ANALYSIS IN FOREST BIOLOGY

J. G. Isebrands and Thomas R. Crow

INTRODUCTION

There is a definite need to acquaint those interested in, yet unfamiliar with principal component analysis (PCA), regarding its terminology, underlying assumptions, practical applications, and literature so that PCA might find more widespread and proper use in data analysis. Although most multivariate textbooks (e.g., Rao 1952, Kendall 1957, and Seal 1964) adequately cover the theoretical aspects of PCA, examples of practical applications with information concerning the interpretation are lacking in the literature. Adding to the confusion for the beginner is the proliferation of matrix notation and the lack of standardization among texts in both notation and terminology.

Our objective is to introduce PCA to the forest biologist who has had an exposure to introductory statistics, and likely applies ANOVA, correlation, and regression routinely, but who has not made the jump to multivariate techniques. Our intent is to demonstrate through detailed examination of two applications the utility of principal component analysis in helping solve research problems in forest biology.

It should be emphasized that PCA is normally not used to test a null-hypothesis or in the estimation and prediction of quantities. Instead, it is an exploratory technique for assessing the dimensions of variability and aiding in the generation of hypotheses to be tested in conjunction with other statistical techniques such as multiple regression (Pearce 1969). Among the many potential uses of PCA in forest biology, those which will receive emphasis in this paper are: (1) reduction of the number of variables by deletion of extraneous variables; (2) ordination of variables as an aid to the interpretation of multivariate data; and, (3) use of PCA in conjunction with regression analysis for the identification of biological variables for further experimentation.

PRINCIPAL COMPONENT ANALYSIS

Historical Development

Principal component analysis (PCA) is certainly nothing new; mathematical statisticians have studied it for years (Hotelling 1933, Rao 1952, Kendall 1957, Anderson 1964, Seal 1964). As research tools the initial development and application of multivariate techniques are rooted in the behavioral sciences. The classical example is Spearman's (1904, 1927) attempt to prove his psychological theory that intellectual performances are a function of a single general mental capacity. The origins of PCA can be traced to variance-maximizing solutions in psychological and educational studies (Hotelling 1933). Recent emphasis given multivariate techniques is associated with the availability of computers to process the extensive calculations associated with the techniques. Almost every computer center now has one or more multivariate packages (e.g., Dixon 1970).

In forest biology, applications of PCA have been relatively few, although there has been a flux of recent publications. J. N. R. Jeffers (1962, 1964, 1965, 1967, 1970, 1972)

has been the greatest proponent of the use of multivariate analysis. Jeffers and Black (1963) applied PCA to 9 lodgepole pine provenances using 19 variables; they concluded that many fewer than 19 variables were needed to discriminate among provenances. Namkoong (1967) also used PCA for an analysis of provenance data in conjunction with regression.

Gessel (1967) recommended the application of PCA to aid in the assessment of the many factors that influence forest productivity, or yield. In an example, eighteen variables were tested against the productive capacity as measured by site index from a series of western hemlock (*Tsuga heterophylla* (Rafn.) Sarg.) stands in Washington State. Four uncorrelated components were found to have a major influence on the patterns of variation in productive capacity (Gessel 1967). Others have also utilized PCA to assess production relationships. Kinloch and Mayhead (1967) investigated the use of PCA to help assess the possibility of using ground vegetation as an indicator of productive potential in forestry. Decourt *et al.* (1969) used PCA and regression analysis with orthogonalized variables to elucidate the relationships between environmental factors and production in Scotch pine (*Pinus sylvestris* L.). PCA was employed by Vallée and Lowry (1972) to classify black spruce (*Picea mariana* (Mill.) B.S.P.) forest types and to help estimate site quality. Auclair and Cottam (1973) employed PCA and multiple regression analysis to assess the influence of environmental factors on the radial growth of black cherry (*Prunus serotina* Ehrh.).

In other forestry related areas, PCA has been used in dendrochronology (Fritts *et al.* 1971, LaMarche and Fritts 1971), palynology (Webb 1973, 1974a, 1974b), and geoecology (Newnham 1968).

Basic Properties

Principal component analysis is an analytical procedure for transforming one set of variates into another set of component variates having the following properties: (1) they are linear functions of the original variates; (2) they are orthogonal, i.e., independent of each other; (3) the total variation among them is equal to the total variation in the original variates, consequently, information concerning differences among the observed variates is not lost in transformation; and, (4) the variance associated with each component decreases in order--the first variate will account for the largest possible proportion of the total variation, the second will account for the largest proportion of the remainder, and so forth.

Bearing these properties in mind, a comparison of PCA to another popular multivariate technique, factor analysis, is appropriate. Within the literature, there is a great deal of conflicting terminology; as a result the distinction between PCA and factor analysis can be confusing. For example, where the term "factor analysis" has been applied to all multivariate procedures dealing with the reduction of dimensionality and identification of common factors, PCA is often presented as a "factor analysis" technique. In other cases, such as the IBM Scientific Subroutine Package, PCA is labeled as "principle component factor analysis."

Two important distinctions exist between PCA and factor analysis:[1]

(1) In factor analysis, p original variates are reduced into m<p uncorrelated "factors" having an uncorrelated residual component; in PCA, p correlated variates are transformed into p uncorrelated variates, not all of which are necessarily significant.

(2) Unlike PCA, factor analysis has the potential for rotating the orthogonal axes that represent "factors" to new oblique positions so that theoretical postulations inherent in a model can be tested.

The first of these distinctions has to do with property No. 3 above. An assumption basic to PCA is that the observed variation is caused by the effects that the underlying (casual) factors have on each of the original variates. PCA, therefore, is a closed model, without regard to random error or variation external to the system (Pearce and Holland 1960); thus, all variation in the original variates is accounted for by the derived variables. In factor analysis, however, only a portion of the total variation is attributed to the m<p transformed variates (this portion is termed the "communalities") and the remaining variance is considered an error variance.

1/ *For details see Kendall (1957), Pearce and Holland (1960), Seal (1964), Cattell (1965), and Pearce (1965).*

Although the factor analysis model may seem more desirable for biological applications, the need to estimate communalities poses a problem because it requires *a priori* knowledge of the system. Initial estimates of communalities often are little better than arbitrary guesses; thus, a series of iterations is necessary before the investigator is satisfied. As a result, the model is developed to fit the data (Kendall 1957). In PCA, however, the process is reversed: one works from the data toward a hypothetical model. Beginning with the observations, the investigator develops a model that reduces the dimensions of variation, which consequently aids in the biological interpretation (Kendall 1957).

It must be emphasized that the variates--principal components--derived using PCA may not have any biological significance. Multivariate techniques must not be considered as a mode for the automatic generation of hypotheses, rather as an initial step in which complex data sets are simplified to make them more amenable to interpretation. Any hypothesis developed using PCA that seems plausible can only be considered subjective until confirmed by existing biological knowledge or additional studies (Pearce 1965).

Terminology and Notation

It is not necessary for the user to understand all aspects of PCA derivation. He must, however, have an overview of its terminology and notation if he is to effectively use PCA.

For example, suppose that x_1, x_2, x_i ... x_p are random variables and that $\underset{\sim}{x}$ is a row vector composed of the x's. From this population, a sample of n independent observations can be drawn so that

$$\underset{\sim}{X} = \begin{bmatrix} x_{11} & \cdots\cdots & x_{1p} \\ \cdot & & \\ \cdot & & \\ x_{n1} & \cdots\cdots & x_{np} \end{bmatrix}$$

where $\underset{\sim}{X}$ is an n x p data matrix of full rank (i.e., independent rows and columns) (Morrison 1967). The variance-covariance matrix of $\underset{\sim}{X}$ is defined equal to S and the correlation matrix of $\underset{\sim}{X}$ is R.

The principal components are defined as linear combinations of the original x_i variables under consideration and are denoted by ζ_i, i = 1 ... p. The ζ_i have the form

$$\zeta_i = a_{i1} x_1 + a_{i2} x_2 + \ldots + a_{ip} x_p$$

$\underset{\sim}{A}$ is defined as a column vector and is referred to as an "eigenvector" (or latent vector) having coefficients a_{ij}. The coefficient subscript i refers to the eigenvector number and the subscript j refers to the original variable (x_i) number.

Each eigenvector ζ_i has a variance associated with it called an "eigenvalue" (or latent root) and is denoted by λ_i, i = 1 ... p.

Geometrically, we have a data scatter of n points in p dimensions and PCA is a rotation of axes such that the total variance of the projections of the points onto the first axis is a maximum (i.e., first principal component). The second axis (second principal component) is chosen orthogonal to the first and accounts for as much as possible of the remaining variance. Each additional axis is also orthogonal and accounts for a maximum portion of the remaining variation (Seal 1964). The linear combinations ζ_i are the length of the projections onto the new axis, and the directional cosines of the projections are the eigenvector coefficients a_{ij}. The variance of the projection is the eigenvalue (λ_i) (Seal 1964, Krzanowski 1971).

The algebraic and matrix derivation of the principal components, ζ_i and their variance is beyond the scope of this paper. However, it has been covered in detail in the literature (Hotelling 1933, Kendall 1957, Anderson 1964, and Morrison 1967).

The following properties of PCA are of importance for interpreting our examples. As stated earlier, the variance of each principal component (ζ_i) is its eigenvalue (λ_i); furthermore, the λ_i values sum to the total variance in the experiment. Therefore,

$$\lambda_1 + \lambda_2 + \lambda_3 + \ldots + \lambda_p = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \ldots + \sigma_p^2$$

This means that the total variation of the derived variates equals the total variation of the observed variates, thus information is not lost by linear transformation.

In addition, the quantity

$$\left(\frac{\lambda_i}{\Sigma\lambda_i}\right)100$$

where $\Sigma\lambda_i$ = trace S (i. e., sum of diagonal elements of the correlation matrix) gives the percentage of the total variance explained by the i^{th} principal component (table 1). The cumulative percent of the total variance is also important because it refers to that portion of the variance "explained" by a particular eigenvector in question plus all previous eigenvectors.

Table 1.--*Eigenvalues, and cumulative percentage of variation associated with the eigenvalues from principal component analysis of 4-year white spruce nursery measurements*

	Eigenvalues (λ)	Cumulative percent of variation
1	10.07	0.530
2	2.25	0.648
3	1.63	0.734
4	1.08	0.791
5	1.05	0.846
6	0.56	0.875
7	0.50	0.902
8	0.49	0.928
9	0.31	0.945
10	0.29	0.950
11	0.23	0.971
12	0.18	0.981
13	0.13	0.988
14	0.09	0.992
15	0.06	0.996
16	0.03	0.998
17	0.03	0.999
18	0.01	0.999
19	0.01	1.000

Operational Sequence

The mathematical operations of PCA are important, but they represent only one aspect of the analysis. The entire spectrum of operational sequence follows:

(1) Selection of preliminary variables; (2) if necessary, transformation of original variables to meet the assumptions - ε's normally and independently distributed with mean 0 and variance σ^2; (3) selection of either the variance-covariance or the correlation matrix and calculation of that matrix; (4) determination of eigenvalues (latent roots) and eigenvectors of the variance-covariance or correlation matrix; and (5) interpretation of derived components.

The first step, variable selection, is extremely important. These variables should be quantitative characters, and preferably be measured on a continuous scale, although many discrete variables (e.g., the number of teeth measured along a leaf margin) adequately approximate continuous variables (Jeffers 1964).

The second step requires deciding whether to transform data, which admittedly can be a subjective decision (Jeffers 1964); in most statistical analyses, the assumption of normality is often neglected. Tests of significance are only meaningful for data that are multivariate-normal in their distribution and transformations may be necessary if normality is not present (Andrews, Gnanadesikan, and Warner 1971). Furthermore, Bartlett's test can be used for testing the homogeneity of variance. However, Jeffers (1964) recommended the use of transformations only when the data severely violate the assumptions, because transformations make the eventual interpretation of PCA more difficult.

The third step calls for another decision whether to use the variance-covariance matrix or the correlation matrix. Normally, if all units are of the same scale (e.g., all units of length), the use of the variance-covariance matrix is recommended. Use of the variance-covariance matrix has the greatest statistical appeal because the sampling theory is less complex than the others (Anderson 1964). However, if the units are mixed (e.g., length, volume, weight), normalization is necessary and the correlation matrix is used. The eigenvalues (variance) associated with an eigenvector from a correlation matrix is a standardized variance. Throughout this paper the correlation matrix is used.

The fourth step involves the linear transformation of p original variates into

p "artificial" variates. This is the mathematical equivalent of determining the eigenvectors and related eigenvalues of a variance-covariance or of a correlation matrix (Jeffers 1964). Conceptually this requires the extraction of common variables (i.e., the eigenvectors) and their variances (i.e., eigenvalues) from the variance-covariance or correlation matrix. A simplistic development of the mathematical derivation of eigenvectors and eigenvalues can be found in Pearce (1969); more complete derivations can be found in any matrix algebra text.

The fifth step involves the interpretation of the derived components. First, a decision has to be made regarding the number of components that have biological significance. There are various criteria to aid in this decision; in general, the elimination of those vectors that do not meet the criteria can be done with conviction. Admittedly, some subjectivity is involved in this process, but this is inherent in all statistical decisions. The next part of the interpretation process is the analysis of the eigenvectors that are deemed significant.

However, one must be cautioned that even after this operational sequence, there still remains the question whether a biological interpretation can be derived from the mathematical artifact. To interpret derived variables, one must be able to relate them to observed variables. To do this there are several accepted ways, which are explained in the example beginning on p. 5.

EXAMPLES OF APPLICATIONS

Discarding Variables

Among the PCA methods for reducing the dimension of a data set by discarding variables, we have found the method of retention outlined by Jolliffe (1972) most useful. The following example demonstrates the use of this method.

In 1958 a range-wide study of white spruce seed sources consisting of 28 provenances originating throughout North America from Alaska to New Brunswick (table 2) was established at our nursery near Rhinelander, Wisconsin. After 4 years' growth, 19 variables were measured on trees from each provenance (table 3). The large number of variables was considered necessary because of the preliminary nature of the study. Information derived from the study was to be used to help determine the usefulness of certain parameters for possible selection indices. These parameters would then be studied further in subsequent experiments. Portions of the data have been published (Nienstaedt 1968, Nienstaedt and Teich 1971).

The data consists of a 19 x 28 matrix suitable for the use of PCA in discarding variables. Bartlett's test of homogeneity indicated that the 19 variances are homogeneous at the 0.05 probability level, and therefore, no transformations are necessary. We continue the discarding procedure by calculating the 19 x 19 correlation matrix from the original data matrix and run PCA on it. See table 1 for the 19 eigenvalues (λ_i) and the cumulative percentage of the total variation "explained" by each.

Next we choose an arbitrary value, called λ_o, which has associated with it at least the cumulative proportion of the total variance that one wishes to "explain" in the analysis. This procedure is somewhat analogous to choosing the probability level that one wishes to operate at in a routine analysis of variance. Therefore, it depends not only upon the experimental material, but also upon experience of the scientist. Jeffers (1964) recommended choosing $\lambda_o = 1$ for biological data. If we choose $\lambda_o = 1$ in this example, we would expect to "explain" approximately 85% of the total variation (table 1) because there are 5 eigenvalues greater than 1.0 and their cumulative percent of variation is 84.6. The subset of λ's that are greater than λ_o is of size p_1; thus, there are also p_1 eigenvectors associated with these p_1 eigenvalues. In this example, $p_1 = 5$.

In examining the eigenvalues after PCA it may be necessary to distinguish between marginal eigenvalues (λ_i). Lawley (1956) showed that the degree of difference between eigenvalues can be measured by the ratio of the geometric mean of the eigenvalues to the arithmetic mean, which is distributed as χ^2. This procedure was outlined by Holland (1969).

The discarding procedure is continued by associating one or more of the variables

Table 2.--*Provenance values of the first 3 eigenvectors used for ordination of the white spruce example*

Source number	Location	Latitude	Longitude	Eigenvector number 1/	2	3
1	South Dakota	44-10	103-65	1/0.86	-3.54	-2.30
2	Montana	46-48	109-31	-6.31	4.64	0.22
3	Manitoba	49-51	99-30	1.26	-3.98	2.39
4	New York	44-23	74-06	10.70	0.11	-0.81
5	Wisconsin	45-41	89-07	8.09	-3.52	0.74
6	Minnesota	47-33	94-09	9.03	1.38	0.44
7	Minnesota	47-33	94-10	9.45	-7.21	0.52
8	New Hampshire	44-51	71-26	8.17	3.36	-0.02
9	Alaska	65-21	144-30	-23.08	0.58	0.77
10	Alaska	63-45	144-53	-13.10	-2.65	0.09
11	Alaska	66-35	145-11	-15.98	-2.08	-1.73
12	Maine	44-50	68-38	8.92	8.91	-1.00
13	Labrador	52-36	56-26	-6.66	5.46	3.00
14	Labrador	53-46	60-05	-5.00	3.21	2.36
15	New Brunswick	47-50	68-21	6.99	5.70	-4.75
16	Quebec	46-32	76-30	8.99	-0.01	0.13
17	Quebec	48-18	71-22	5.42	3.66	-0.52
18	Ontario	48-00	81-00	16.53	0.27	3.90
19	Ontario	45-44	76-51	12.38	-5.67	-0.20
20	Manitoba	54-39	101-36	-3.67	-3.16	-3.73
21	Saskatchewan	59-19	105-59	-8.50	-1.76	-1.70
22	Yukon	60-49	105-35	-18.51	-1.21	1.73
23	Minnesota	47-33	94-08	5.59	-2.83	-1.51
24	Michigan	44-30	83-45	4.46	-0.52	2.09
25	British Columbia	54-00	123-00	-2.80	0.96	-1.07
26	Manitoba	56-56	92-51	-16.12	0.62	-0.08
27	Ontario	52-15	81-40	-0.98	-0.47	1.24
28	Ontario	48-30	89-30	3.88	-0.25	-0.21

1/ Calculation procedure for each value:

$$\sum_{1}^{j} \{ ((x_{ij} - \bar{x}_j) / s_j) \cdot a_{ij} \}$$

under consideration with each of the p_1 eigenvectors mentioned above (Spurrell 1963, Beale *et al.* 1967, Namkoong 1967). This involves choosing the coefficient or coefficients having the highest absolute value in each eigenvector starting with the first eigenvector. Table 3 shows coefficients for the five eigenvectors (components) associated with the first five eigenvalues ($\lambda_o = 1$; $p_i = 5$). The variables circled in table 3 should be retained. In our example, four coefficients in eigenvector 1, which accounts for 53 percent of the total variation, are candidates for having the highest absolute value in the vector because they are approximately equal (0.300): height (x_1); diameter (x_2); branch length (x_4); and bud length (x_7). Therefore, these variables are retained.

Next consider eigenvector 2. The largest coefficient in this vector is 0.410 and is associated with bud color (x_6); therefore, bud color is retained. In eigenvector 3 the highest coefficient is associated with number of adaxial stomata (x_{13}); in eigenvector 4 with incidence of second flushing (x_{18}); and in eigenvector 5 with needle color (x_{11}). These variables are also retained.

The sign of the largest coefficient can be either positive or negative because the highest coefficient is chosen on

Table 3.--*Variables measured from 4-year-old white spruce provenances and the first five eigenvectors from the principal component analysis*

List of variables	Eigenvectors (A_i)				
	1	2	3	4	5
x_1 Height (in.)	0.303	-0.057	0.053	-0.088	0.137
x_2 Diameter (mm)	0.305	0.034	-0.013	-0.010	-0.051
x_3 No. of branches in top whorl	0.236	0.002	0.087	-0.248	-0.431
x_4 Branch length in top whorl (mm)	0.298	0.004	-0.032	0.009	0.177
x_5 Bud shape	0.176	0.376	-0.195	-0.085	-0.065
x_6 Bud color	0.164	0.410	0.044	0.211	0.215
x_7 Bud length (mm)	0.302	0.084	-0.048	-0.003	-0.126
x_8 Needle length (mm)	0.195	-0.312	-0.271	-0.257	0.021
x_9 Needle shape	-0.232	-0.009	0.314	-0.234	-0.166
x_{10} Needle rigidity	0.286	-0.116	-0.014	-0.006	-0.016
x_{11} Needle color	-0.045	0.268	0.255	-0.429	0.581
x_{12} Needle curvature	0.195	0.181	0.424	0.288	0.000
x_{13} Stomata (adaxial)	-0.006	-0.269	0.665	0.025	-0.060
x_{14} Stomata (abaxial)	0.216	-0.367	0.190	-0.154	0.199
x_{15} Needle serrulation	-0.235	-0.196	-0.197	0.029	0.349
x_{16} Branch surface	-0.266	0.000	0.001	-0.008	-0.346
x_{17} Sterigmata length (mm)	0.225	-0.276	-0.089	-0.269	-0.018
x_{18} Second flushing	0.101	-0.342	-0.031	0.620	0.154
x_{19} Forking	-0.273	-0.159	-0.061	-0.092	0.144

the basis of absolute value. Furthermore, the variables associated with the eigenvectors cannot be ones that are already associated with an earlier vector. When this occurs, the second largest coefficient is chosen. Questions have been raised about using this approach. However, Brown, Douglas, and Wilson (1971) showed that the coefficients of the original variables in the eigenvectors are not affected by the intercorrelation of the x's; therefore, the largest coefficient approach is valid.

The 8 variables we have retained--height, diameter, branch length, bud color, number of adaxial stomata, needle color, and amount of second flushing--are those to be considered in further experimentation. All others are discarded. This means that in our example, x_3, x_5, x_8, x_9, x_{10}, x_{12}, x_{14}, x_{15}, x_{16}, x_{17}, and x_{19} are rejected (table 3).

In this particular example, we happen to have hindsight as to the nature of the variables. Furthermore, those retained for further consideration appear to apply biologically. The first eigenvector can be considered a vector of size, because bud length at nursery age, height, diameter, and branch length all are important indicators of growth. This indicates that bud length and branch length may be as important as the traditional measurements--height and diameter--for distinguishing nursery age provenances. The retention of bud color and needle color also seems logical because both are important distinguishing characteristics of nursery-age white spruce. Needle color is particularly important in distinguishing the western provenances where introgression with Engelmann spruce has occurred. Similarly, the retention of second flushing is logical because it is an indicator of the latitude of the origin of white spruce, which can be related to the number of growing days. Second flushing is, therefore, an indicator of growth potential of white spruce.

The retention of number of adaxial stomata indicates that needle anatomy

data may be useful. However, the utility of anatomical data as selection indices must be weighed against the time and expense of collecting such data.

Ordinating Groups of Variables

Ordination, the ordering of units within a multidimensional space, has had widespread application in many other ecological studies and has potential for application in many other biological areas. The use of ordination is consistent with the desire to simplify and code a diversity of information so the underlying patterns of variability within a large data set can be more easily grasped.

When the entities (in this example, the 28 provenances of white spruce) are cast into this multidimensional hyperspace using the eigenvectors and associated components as axes (X-axis corresponds to the first eigenvector, Y-axis to the second eigenvector, etc.), the distance among these points is proportional to the degree of dissimilarity in terms of a set of variates (i.e., properties, measured parameters, characters). Thus an ordination has occurred. Furthermore, if discrete subpopulations with some degree of biological integrity can be defined, a classification can be obtained.

Using the white spruce data we obtained a PCA ordination as demonstrated by Jeffers and Black (1963). For each original variate of a given provenance, a standardized variable (which is the difference from the mean of all provenances divided by the standard deviation) was obtained and multiplied by the appropriate eigenvector coefficient found in table 3. Summation over all the variables

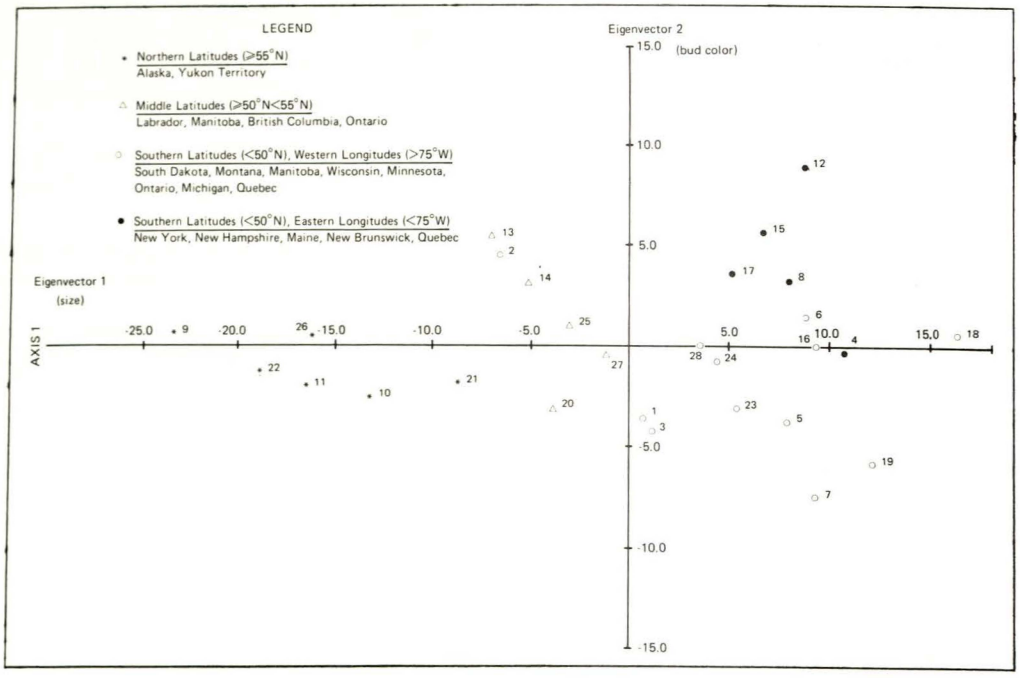


Figure 1.--*Ordination of white spruce provenances along two axes corresponding to eigenvector 1 (size) and eigenvector 2 (bud color). The position of each point is determined by the provenance values given in Table 2. The number associated with each plotted point is the source location number listed in Table 2. From a visual perspective, this figure can be considered a two-dimensional "side view" of an ellipsoid in three-dimensional space.*

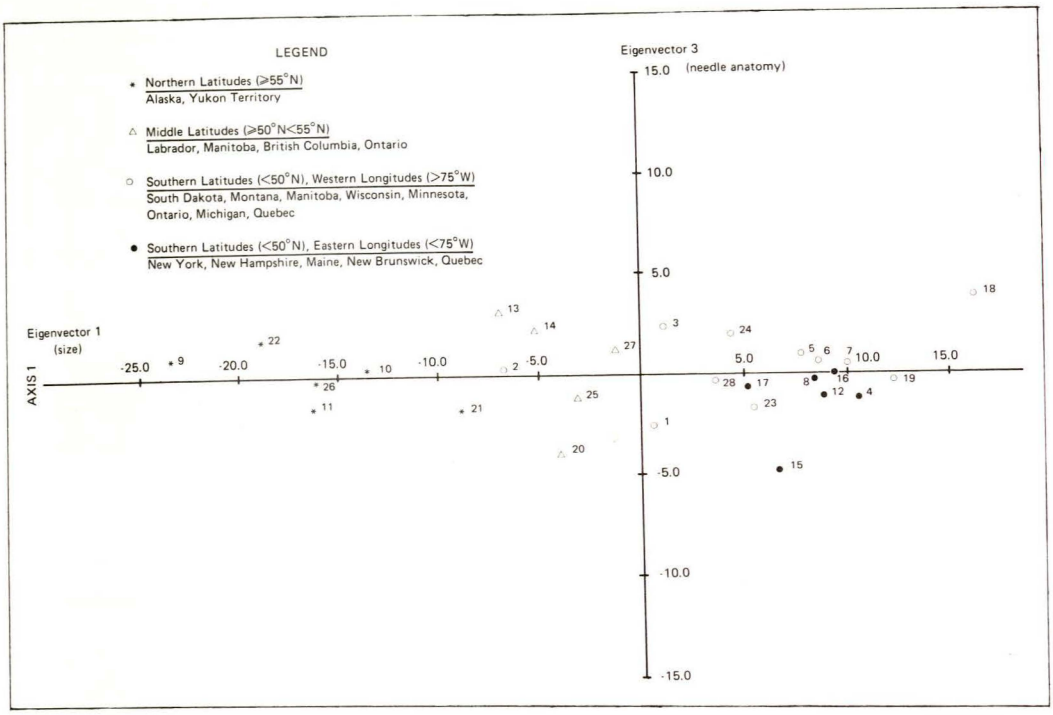


Figure 2.--*Ordination of white spruce provenances along eigenvectors 1 (size) and 3 (needle anatomy). This figure can be considered a "top view" of the ellipsoid.*

in the eigenvector then provides the numerical value for each provenance found in table 2 and plotted in figures 1, 2, and 3.

The most striking characteristic of the ordination of the white spruce data is the elongated form of the hypersolid. This confirms the importance of the first component (the size factor). This also suggests that the underlying dimensions of variability can be represented by far fewer than 19 variables with little or no loss of information.

The ordination is largely dependent on the first component--size--and the order of points in figure 1 corresponds to changes in latitude and corresponding elements such as length of growing season, temperature regime, and photoperiod; all affect the expression of the genotype (i.e., the phenotype) of a source and thus affect performance as measured by size in a provenance study. From left to right in figure 1, the first group of points is made up of provenances from the highest latitudes and the northwestern portions of the white spruce range; the second group is from the middle latitudes, and those to the right of the second axis (eigenvector 2) are from the lower latitudes and the southeastern portion of the range. In terms of size the poorest performers are on the left in figure 1, progressing in an orderly fashion along the first axis to the best performers on the right.

The second component--bud coloration--is important in discriminating among the eastern seed sources. Note the vertical spread in the points of the right quadrants (fig. 1). The ordering of points suggests the possibility of clinal variation in bud coloration along the longitudes. The points in the upper right (++) quadrant in figure 1 are seed sources from the eastern longitudes within the southeastern portion of the white spruce range; those in the lower right (+-)

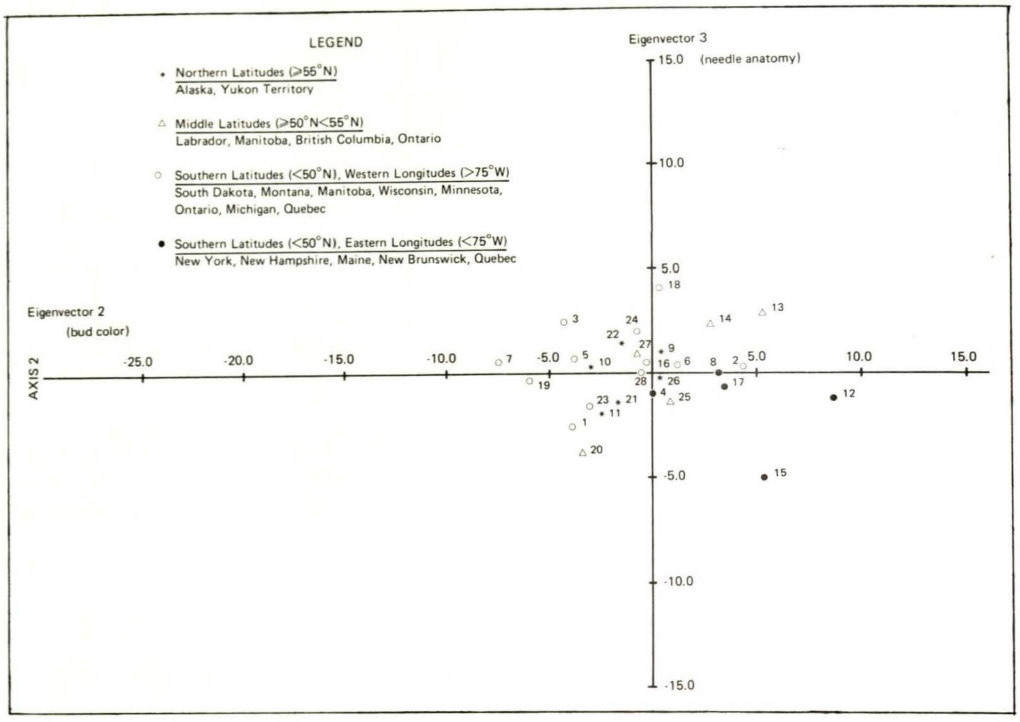


Figure 3.--*Ordination of white spruce provenances along eigenvectors 2 (bud color) and 3 (needle anatomy). This figure can be considered an "end view" of the ellipsoid.*

quadrant are from the western longitudes within this sub-group.

The third component--needle anatomy--in figure 2 or 3 provides little discrimination among points. This gives credence to the fact that few orthogonal variables were measured.

In a review of the systematics of white spruce, Nienstaedt and Teich (1971) cite evidence for the division of the species into eastern and western populations. Needle characteristics such as color and length were cited as major contributors to the east-west variation pattern. The fact that the present white spruce population has evolved from populations that survived both the Illinoian and Wisconsin glaciations in widely separated refugia is also given as supporting evidence. Studies of monterpenes in cortical samples (Wilkinson *et al.* 1971) and DNA content per cell (Miksche 1968) also support the contention of two distinct populations.

As recognized by Nienstaedt and Teich (1971) and supported by the results of our PCA, the demarcation of white spruce into two populations would appear to be an oversimplification; however, the phytochemical characters cited above were not included in this analysis and would no doubt add new dimensions of discrimination if included. In support of the hypothesis of separate populations, the response to the second component, bud coloration, did differ in the western and eastern seed sources. However, no east-west variation pattern is evident in the needle anatomy (specifically, the number of upper surface stomata), the third principal component.

The preliminary nature of the white spruce study and the stated objective of measuring a large number of variables to assess their value as selection indices presents an ideal situation for the application of PCA. A comparison between the results of our analyses and those based on

analysis of variance (table 4) by Nienstaedt and Teich (1971) illustrates the value of PCA. Their analysis of variance shows that all 19 characteristics except needle color and second flushing were significantly different among provenances at the 0.01 level (table 4); thus, the ANOVA does not provide insight into the underlying dimensions of variability, nor does it provide guidance for the selection of variables suitable for emphasis in further studies.

Table 4.--*Analysis of variance of 19 characteristics measured on nursery grown white spruce representing 28 provenances from the entire range of the species* 1/

Variable	F value
Height (in.)	15.61*
Diameter (mm)	30.75*
No. of branches	4.35*
Branch length	10.63*
Shape of bud	2.77*
Bud color	5.67*
Length of buds	15.95*
Needle length	5.40*
Cross section of needle	2.94*
Needle rigidity	2.55*
Needle color	n.s.
Needle curvature	2.42*
No. of stomata upper	2.72*
No. of stomata below	4.59*
Needle serrulation	6.08*
Branch surface	8.05*
Sterigmata length	4.68*
Secondary bud flushing	n.s.
Forking	3.78*

1/ From Nienstaedt and Teich (1971)
* Significant at the 1 percent level

It is noteworthy that the two non-significant variables in the ANOVA were identified by PCA as important orthogonal variables worthy of further consideration. Substantial variation is known to occur in these two characteristics. It is also possible that the strong interdependence of these variables resulted in a non-significant interpretation among provenances in the non-orthogonal ANOVA.

Principal component ordination is one of a multitude of ordination techniques, but is not necessarily the most effective. In the first comparison of principle component ordination to other techniques that numerically approximate multivariate analysis (e.g., Bray and Curtis 1957, Swan, Dix, and Wehrhahn 1969), PCA was found to be superior (Austin and Orloci 1966, Orloci 1966). But subsequent studies by other authors have obtained the contrary conclusion (Bannister 1968, Austin and Noy-Meir 1971, Gauch and Whittaker 1972, Whittaker and Gauch 1972, Gauch 1973).

It is evident from these evaluations that like any mathematical technique, ordination is most effective when the user is aware of the limitations as well as the capabilities of the particular technique (Gauch 1973). Ordination is a linear mapping technique and if the parameters under study respond to an experimental stimulus in a nonlinear fashion (i.e., a non-monotonic performance or response), the representation of the parameter/stimulus relation in a multidimensional space (ordination) may be distorted. For example, in the ecological sphere, the response of vegetation to environmental gradients is highly nonlinear. In such a case, if one does not recognize the discrepancy between the linear assumption of the ordination methods and the nonlinear response by the biological system, evaluation of vegetational patterns as influenced by environmental factors can lead to spurious conclusions.

Multiplotting

It is often desirable or necessary to visualize the results in as many dimensions as possible, but plotting of multivariate data is limited by human perception. To circumvent this problem, contouring or the addition of symbols can be used to extend the number of axes on a two-dimensional plot. However, such plots lack precision and soon become difficult to interpret with the additional clutter. Physical models or 3-D plots from stereo equipment can be used for interpretation of multivariate data, but n-dimensional data still cannot be perceived.

To solve this problem, Andrews (1972) suggested that one should map points into a function and then plot the function. A function can be infinite in its dimensions

and still be easily visualized in two-dimensional space. This allows interpretation in more than three dimensions. The function proposed by Andrews (1972) follows:

$$f_\zeta(t) = \zeta_1/\sqrt{2} + \zeta_2 \sin t + \zeta_3 \cos t + \zeta_4 \sin 2t + \zeta_5 \cos 2t$$

For each point the function is defined and then plotted over the range

$$-\pi < t < \pi \qquad \zeta_i = \sum_1^j x_{ij} a_{ij}$$

Thus a set of points is transformed into a set of lines that has many appealing properties.

For example, the functional mean corresponds to the mean of the observations themselves: that is, if $\bar{x}$ is the mean of n-multivariate observations, the function corresponding to $\bar{x}$ will appear as an average on the plots. Such a set of lines also preserves distances. If plotted functions are close together for all values of t, the corresponding points are close together in n-dimensional space and a band of functions represents a cluster of data points. If a group of functions are close together for only one value of t, the corresponding points are close in the direction defined by the corresponding projection in one-dimensional space. Therefore, even with n-dimensions groups of points can be identified.

This function also preserves variances. Therefore, tests of significance and confidence intervals can be constructed at particular values of t because the variance of $f_\zeta(t)$ is known.

The major advantage of multiplotting is not the establishment of variation patterns based on a single orthogonal character or even several characters, but the discrimination among populations based on the integral of characters. For example, Jeffers (1972) distinguishes a number of birch species by multiplotting five components of 13 leaf characters. In his example, several birch hybrids are evident from their intermediate position on the plots. In addition, Andrews (1972) demonstrates how biological data can be misinterpreted when only two-dimensional point plotting is used.

Principal Components in Conjunction With Regression Analysis

Many forest biologists often wish to build a model to predict a dependent variable (Y) from a complex set of interrelated independent variables (x's). When selecting their variables, they often are faced with a dilemma. They not only want to include the variables that are most influential in controlling the systems, but also enough variables to obtain a reasonable fit, so their models are useful for predictive purposes (Draper and Smith 1966). However, large numbers of variables increase the complexity of a model tremendously (Goodall 1972). Therefore, they may wish to choose a method of variable selection when building a preliminary model.

If adequate degrees of freedom are available, many would prefer to perform preliminary analysis and model building on one-half of the data and then test the model on the other half to validate the model. However, others feel that this is unnecessary in preliminary experiments because data will be gathered subsequently to test the model and to update it. We chose the latter approach in the following example.

Many methods are now available for selecting variables to use in a regression equation: these include the all regressions approach, backward elimination, forward selection, stepwise regression, and several combined techniques. Several of these methods do not give satisfactory results when the intercorrelation between the x's is high (Draper and Smith 1966). This is because under conditions of normality, the higher the correlations between variables, the less orthogonal the data will be (Draper and Smith 1969). Furthermore, the selection methods don't necessarily help us select the best equation; but usually they will allow us to find an acceptable one.

Another problem that must be considered is that some of these methods require repeated tests of significance; therefore, they are based on conditional decisions (i.e., one test

influenced by the previous test). In this case, one may be operating at a different probability level than expected. Little consideration usually is given to the consequences of such conditional tests (Kennedy and Bancroft 1971).

Principal component analysis also can be used in conjunction with multiple regression to select variables for a regression equation. Various approaches have been reported by Kendall (1957), Ahamad (1967), Beale *et al.* (1967), Jeffers (1967), Spurrell (1963), Cox (1968), and Decourt *et al.* (1969).

Kendall used the standard multiple regression model:

$$Y = b_o + b_1 \zeta_1 + b_2 \zeta_2 + \ldots + b_p \zeta_p + \varepsilon \quad (1)$$

in which he substituted ζ_i's for x's where ζ_i = i^{th} principal component from the set of variables.

By applying the principle of least squares, the estimates of the b's are obtained by solving the set of normal equations. In this case the coefficients

$$b_i = \frac{\Sigma y \zeta_i}{\Sigma \zeta_i^2} \quad (2)$$

as in orthogonal polynomials (Anderson and Houseman 1942). Furthermore, the reduction due to fitting the regression on the ζ_i's is $b_i \Sigma y \zeta_i$, which is also equal to $\lambda_i b_i^2$. Solving this equality for b_i, we obtain

$$b_i = \frac{\Sigma y \zeta_i}{\lambda_i} \quad (3)$$

which can be used in evaluating the contribution of the original variables as follows.

In Kendall's procedure, PCA is run on the correlation matrix of the original set of variables. Then the b_i's are found by solving equation (3). When solving for b_1, each eigenvector coefficient in eigenvector 1 is multiplied by the correlation coefficient of that x and Y and then summed for the eigenvector. This value is then divided by its eigenvalue (λ_i) (i.e., $b_1 = \frac{\Sigma y \zeta_1}{\lambda_1}$).

All eigenvectors having eigenvalues near zero are neglected because they contribute little to the total variance. The total variance is calculated for each b_i by solving

$$\sigma^2_{b_i} = \lambda_i b_i^2 \quad (4)$$

To evaluate the contribution of the x's, Kendall substituted the b_i's and the ζ_i's in terms of standardized x's into the original equation (1), which produced an equation of coefficients and standardized x's. The b_i's reflect both the sign and sizes of each x variable's contribution.

This approach is most useful when the number of variables is small. However, the problem with this approach is that when the number of variables is large it is often difficult to interpret the results in terms of the individual variables that are embedded in the linear combination (eigenvectors). There also is often some question as to whether the dimension of the problem is truly reduced because the components have contributions from all the x's. Therefore, we believe this approach should be used only when the experimenter can assign biological meaning to certain significant components (eigenvectors), or when the number of variables is small.

Cox (1968) advocated the use of PCA in preliminary experiments to suggest regressor variables. In his method the principal components themselves are not used in the regression equation as in Kendall's procedure. Rather, Cox used simple combinations of variables having physical meaning. We have chosen Cox's approach to illustrate our second example.

In this example we have used the data of Larson (1967). He measured 12 growth variables on trees from 10 red pine seed sources

grown under various conditions in controlled growth rooms (table 5).

After a Bartlett's test indicated that the data had homogeneity of variance, multiple regression analysis was run of the 10 independent variables upon volume increment. The ANOV table for regression is shown in table 6. Note the regression was significant and the $R^2 = 0.98$. It should be emphasized, however, that Larson (1967) did not relate his variables to volume increment as we have done, nor did he suggest this relationship. Rather, we have arbitrarily picked volume increment as the dependent variable for purposes of illustration.

Table 5.--*Selected tree growth measurements and first four eigenvectors of principal component analysis from 10 red pine provenances grown in growth rooms* 1/

List of variables		Eigenvectors (A_i) 1	2	3	4
x_1	Height (cm)	0.351	-0.031	-0.381	-0.163
x_2	Needle length, 1962 (cm)	-0.310	-0.111	0.370	0.576
x_3	Needle weight, 1962 (gm)	0.382	-0.020	0.045	-0.112
x_4	Needle weight, 1961 (gm)	0.384	0.040	0.065	-0.239
x_5	Total ring width (mm)	0.319	0.368	0.250	0.254
x_6	Earlywood width (mm)	0.371	0.053	0.268	0.213
x_7	Latewood width (mm)	0.124	0.711	0.054	0.145
x_8	Latewood percent (%)	-0.275	0.501	-0.234	-0.063
x_9	Specific gravity	-0.343	0.281	-0.263	-0.137
x_{10}	Cell wall thickness (μ)	0.199	-0.111	-0.671	0.647
Y_1	Volume increment (mm^3)				

1/ Adapted from Larson (1967); in the original study, the author made no attempt to relate the independent variables listed to volume increment.

Table 6.--*ANOV for regression of 10 selected red pine growth measurements (x's) on volume increment (Y)* (R^2=0.98)

Source	df	S.S.	M.S.	F
Regression	10	1977630.25	197763.03	69.37*
Deviations	29	82673.52	2850.81	
TOTAL	39	2060303.77		

* Denotes significance at the 0.01 probability level.

Next, a principal component analysis was run on the 10 x 10 correlation matrix of the independent variables (x's). The 10 eigenvalues (λ's) and the cummulative percentage of the total variation associated with each are shown in table 7. Kendall (1957) has shown that when collinearities exist in the x's (i.e., some λ's near zero) no reliance can be put on the individual coefficients in regression equations which include all the variables. Note that collinearities exist in our example since eigenvalues 5 through 10 are near zero (table 7). Consequently, for each eigenvalue near zero, one variable can be expressed in terms of the other variables, and, therefore, the number of variables can be reduced (Kendall 1957, Seal 1964).

Table 7.--*Eigenvalues and cumulative percentage of variation associated with the eigenvalues from principal component analysis of red pine provenance grown in the growth room*

	Eigenvalues	Cumulative percent of variation
1.	6.45	0.645
2.	1.75	0.820
3.	1.11	0.932
4.	0.53	0.984
5.	0.09	0.993
6.	0.04	0.998
7.	0.02	0.999
8.	0.00	1.000
9.	0.00	1.000
10.	0.00	1.000

Selection of the variables to retain for regression is done in the same manner as in our white spruce example, that is, choosing the variables having the largest coefficient absolute value in the most significant eigenvectors.

Several regression equations may be picked using this method which on the surface may appear to be very different. However, for predictive purposes the equations often are equally effective (i.e., have a large R^2 and a good fit) (Kendall 1957).

In our PCA run it seemed likely that at least six or seven variables might be eliminated from the analysis and still have a regression equation that "explained" a large portion of the total sums of squares. Six of the eigenvalues were near zero (eigenvalue 5 through 10) (table 7); therefore, these 6 variables were no doubt interrelated with 4 or 5 more important variables.

The first four eigenvectors and their respective coefficients are shown in table 5. When λ_0 is chosen equal to 1.0, as in the white spruce example, these four eigenvectors "explain" 98.4 percent of the total variation in the independent variables; similarly, the first three eigenvectors "explain" 93.2 percent of the variation (table 7).

The first eigenvector from table 5 "explained" 64.5 percent of the total variation in the independent variables (x's). It has two coefficients that qualify for the largest absolute value because they are approximately the same equal magnitude, needle weight, 1962 (x_3), and needle weight, 1961 (x_4) (circled in table 5). Similarly, in eigenvector 2, the coefficient having the largest absolute value is latewood width (x_7). In both eigenvectors 3 and 4, the largest coefficient is cell wall thickness (x_{10}).

When one encounters a variable in an eigenvector with a coefficient of largest absolute value that has already been associated with a previous eigenvector, then the next largest coefficient in the eigenvector is chosen. Therefore, in eigenvector 4, inasmuch as cell wall thickness is already associated with eigenvector 3, the next largest coefficient in eigenvector 4 is needle length (x_2); therefore, needle length is chosen.

The five variables chosen from the first 4 eigenvectors are the ones to consider for regression analysis. The ANOV for regression of these five variables on volume increment is shown in table 8. The regression was significant and that R^2 = 0.92, and examination of the residuals indicates a good fit. This indicated that 5 independent variables "explained" nearly as much of the total variation in volume increment as did all 10 independent variables (table 6).

The examination of eigenvectors in search for the largest coefficient must be done with caution especially when two variables have a correlation coefficient (r)

Table 8.--*ANOV for regression of needle length, needle weight (1962), needle weight (1961), latewood width, and cell wall thickness on volume increment (Y)* (R^2=0.92)

Source	df	S.S.	M.S.	F
Regression	5	1902309.43	380461.89	81.9*
Deviations	34	157996.60	4646.95	
TOTAL	29	2060306.03		

* Denotes significance at the 0.01 probability level.

near ± 1. When this occurs, scientific insight must be often used in favor of a set of mathematical techniques.

For example, in our case, needle weight for 1962 and needle weight for 1961 are highly correlated (near ± 1); therefore, only one need be included in the analysis.

Because cell wall thickness is difficult and expensive to measure and is associated with the third and fourth eigenvectors, one might be tempted to look at a regression of needle weight, latewood width, and needle length on volume increment and leave out cell wall thickness. Table 10 shows the ANOV for this regression. The regression was significant and R^2 = 0.90.

It appears that the variables chosen by this analysis "explain" a high portion of the regression sums of squares for volume increment. Although needle weight, needle length, and latewood width have some indirect biological integrity as predictors of volume, the main purpose for including this example was to illustrate the various aspects of variable selection that an experimenter may use for further experimentation.

Table 9.--*ANOV for regression of needle length, needle weight (1962), latewood width and cell wall thickness on volume increment (Y)* (R^2=0.91)

Source	df	S.S.	M.S.	F
Regression	4	1868894.50	467223.62	85.4*
Deviations	35	191411.53	5468.90	
TOTAL	39	2060306.03		

* Denotes significance at the 0.01 probability level.

Needle weight for 1962 was retained because of ease of measurement and needle weight for 1961 was dropped. The ANOV for regression of the four other variables is shown in table 9. By dropping the variable needle weight for 1961, the R^2 was reduced from 0.92 to 0.91; therefore, little was lost. Examination of the residuals also indicated a good fit.

Table 10.--*ANOV for regression of needle length, needle weight (1962), and latewood width on volume increment (Y)* (R^2=0.90)

Source	df	S.S.	M.S.	F
Regression	3	1859829.03	619943.01	111.3*
Deviations	36	200477.00	5568.80	
TOTAL	39	2060306.03		

* Denotes significance at the 0.01 probability level.

LITERATURE CITED

Ahamad, B. 1967. An analysis of crimes by the method of principal components. Appl. Stat. 16:17-35.

Anderson, R., and E. Houseman. 1942. Tables or orthogonal polynomial values extended to N = 104. Iowa Agric. Home Econ. Exp. Stn. Res. Bull. 297. p. 594-672.

Anderson, T. W. 1964. Introduction to multivariate statistical analysis. p. 272-287. John Wiley and Sons, New York.

Andrews, D.F. 1972. Plots of high-dimensional data. Biometrics 28:125-136.

Andrews, D.F., R. Gnanadesikan, and J. L. Warner. 1971. Transformations of multivariate data. Biometrics 27:825-840.

Auclair, A. N., and G. Cottam. 1973. Multivariate analysis of radial growth of black cherry (*Prunus serotina* Ehrh). In southern Wisconsin oak forests. Am. Midl. Nat. 89:408-425.

Austin, M. P., and I. Noy-Meir. 1971. The problem of non-linearity in ordination: experiments with two-gradient models. J. Ecol. 59:763-773.

Austin, M.P., and L. Orloci. 1966. Geometric models in ecology. II. An evaluation of some ordination techniques. J. Ecol. 54:217-227.

Bannister, P. 1968. An evaluation of some procedures used in simple ordinations. J. Ecol. 56:27-34.

Beale, E., M. Kendall, and D. Mann. 1967. The discarding of variables in multivariate analysis. Biometrika 54:357-366.

Bray, J. R., and J. T. Curtis. 1957. An ordination of the upland forest communities of southern Wisconsin. Ecol. Monogr. 27:325-349.

Brown, D., A. Douglas, and A. Wilson. 1971. On the interpretation of principal components. IUFRO (Sect. 25), Newsl. 9, p. 2-5.

Cattell, R. 1965. Factor analysis: An introduction to essentials. Biometrics 21:190-251.

Cox, D. R. 1968. Notes on some aspects of regression analysis. J. Roy. Stat. Soc. Serv. A 131 (Pt. 3):265-279.

Decourt, N., M. Godron, F. Romane, and R. Tomassone. 1969. Comparison of various methods of statistical analysis of the relation between environment and production of the Scotch pine in Sologne. Ann. Sci. For. 26:413-443.

Dixon, W. J. 1970. BMD: Biomedical computer programs. Univ. Calif. Pub. Auto. Comp. 2:150-168.

Draper, N., and H. Smith. 1966. Applied Regression Analysis. 407 p. John Wiley and Sons, Inc., New York.

Draper, N., and H. Smith. 1969. Methods for selecting variables from a given set of variables for regression analysis. Bull. Int. Stat. Inst. 43:7-15.

Fritts, H., T. Blasing, B. Hayden, and J. Kutzbach. 1971. Multivariate techniques for specifying tree-growth and climate relationships and for reconstructing anomalies in paleoclimate. J. Appl. Meteorol. 10:845-864.

Gauch, H.G. 1973. A quantitative evaluation of the Bray-Curtis ordination. Ecology 54:829-836.

Gauch, H.G., and R.H. Whittaker. 1972. Comparison of ordination techniques. Ecology 53:446-451.

Gessel, S.P. 1967. Concepts of forest productivity. XIV. IUFRO Congr. Proc. (Sec. 21.) p. 36-50. Munich.

Goodall, D.W. 1972. Building and testing ecosystem models. In Jeffers, J. N. R. 1972. Mathematical models in ecology. Blackwell Scientific, Oxford.

Holland, D. A. 1969. Component analysis: An aid to the interpretations of data. Exp. Agric. 5:151-164.

Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. J. Ed. Psych. 24:417-441; 498-520.

Jeffers, J. N. R. 1962. Principal component analysis of designed experiments. Statistician 12:230-242.

Jeffers, J. N. R. 1964. Principal component analysis in taxonomic research. Great Britain For. Comm. Stat. Sec. Pap. 83:1-21.

Jeffers, J. N. R. 1965. Correspondence on paper by Draper in Statistician 14: 311-318. Statistician 15:207-208.

Jeffers, J. N. R. 1967. Two case studies in the application of principal component analysis. Appl. Stat. 16:225-236.

Jeffers, J. N. R. 1970. A woodland research strategy based on mathematics and computers. Commonwealth For. Rev. 49:275-282.

Jeffers, J. N. R. 1972. Plotting of multidimensional data. Merlewood Research and Development Pap. 35. 7p.

Jeffers, J. N. R., and J. Black. 1963. An analysis of the variability in *Pinus contorta*. Forestry 36:199-218.

Jolliffe, I. T. 1972. Discarding variables in a principal component analysis. I. Artificial data. Appl. Stat. 21:160-173.

Kendall, M. G. 1957. A course in multivariate analysis. 185 p. Hafner Publishing Co., New York.

Kennedy, W. J., and T. A. Bancroft. 1971. Model building for prediction in regression based upon repeated significance tests. Ann. Math. Stat. 42:1273-1284.

Kinloch, D., and G. J. Mayhead. 1967. Is there a place for ground vegetation assessments in site productivity predictions? XIV IUFRO-Congr. Proc. (Sec. 21) p. 246-260. Munich.

Krzanowski, W. J. 1971. The algebraic basis of classical multivariate methods. Statistician 20:51-61.

LaMarche, V. C., Jr., and H. C. Fritts. 1971. Anomaly patterns of climate over the western United States, 1700-1930, derived from principal component analysis of tree-ring data. Monthly Weather Rev. 99:138-142.

Larson, P. R. 1967. Effects of temperature on the growth and wood formation of ten *Pinus resinosa* sources. Silvae Genet. 16:58-65.

Lawley, D. N. 1956. Tests of significance for the latent roots of covariance and correlation matricies. Biometrika 43:128-136.

Miksche, J. P. 1968. Quantitative study of intraspecific variation of DNA per cell in *Picea glauca* and *Pinus banksiana*. Can. J. Genet. Cytol. 10:590-600.

Morrison, D. F. 1967. Multivariate statistical methods. p. 221-258. McGraw-Hill Inc., New York.

Namkoong, G. 1967. Multivariate methods for multiple regression provenance analysis. XIV IUFRO-Cong. Proc. (Sec. 22.) p. 308-318. Munich.

Newnham, R. M. 1968. A classification of climate by principal component analysis and its relationship to tree species distribution. For. Sci. 14:254-264.

Nienstaedt, H. 1968. White spruce seed source variation and adaptation to 14 planting sites in Northeastern U. S. and Canada. 11th Meet. Comm. For. Tree Breeding, Quebec. Part 2., p. 183-194.

Nienstaedt, H., and A. Teich. 1971. Genetics of white spruce. USDA For. Serv. Res. Pap. WO-15, Washington, D.C.

Orloci, L. 1966. Geometric models in ecology. I. The theory and application of some ordination techniques. J. Ecol. 54:193-215.

Pearce, S. C. 1965. Biological statistics: An introduction. p. 180-196. McGraw-Hill, New York.

Pearce, S. C. 1969. Multivariate techniques of use in biological research. Expl. Agric. 5:67-77.

Pearce, S. C., and D. A. Holland. 1960. Some applications of multivariate methods in botany. Appl. Stat. 9:1-7.

Rao, C. R. 1952. Advanced Statistical Methods in Biometric Research. 390 p. John Wiley and Sons. New York

Seal, H. 1964. Multivariate statistical analysis for biologists. p. 101-122. Methuen and Co., Ltd., London.

Spearman, C. 1904. General intelligence objectively determined and measured. Am. J. Psych. 15:201-293.

Spearman, C. 1927. The abilities of man. 415, 33 p. Macmillan: New York

Spurrell, D. J. 1963. Some metallurgical applications of principal components. Appl. Stat. 12:180-188.

Swan, J. M. A., R. L. Dix, and C. F. Wehrhahn. 1969. An ordination technique based on the best possible stand-defined axes and its application to vegetational analysis. Ecology 50: 206-212.

Vallée, G., and G.L. Lowry. 1972. Application of multiple regression and principal component analysis to growth production and phytosocialogical studies of black spruce stands. Quebec Dept. of Lands and For., Res. Pap. 7, 101 p.

Webb, T. III. 1973. A comparison of modern and presettlement pollen from southern Michigan (U.S.A.). Review of Paleobotany and Palynology 16:137-156.

Webb, T. III. 1974a. A vegetational history from northern Wisconsin: Evidence from modern and fossil pollen. Am. Midl. Nat. 92:10-34.

Webb, T. III. 1974b. Corresponding patterns of pollen and vegetation in lower Michigan: A comparison of quantitative data. Ecology 55:17-28.

Whittaker, R. H., and H. G. Gauch. 1972. Evaluation of ordination techniques. In R. H. Whittaker (ed.) Handbook of Vegetation Science, 5:287-321, Ordination and Classification of Communities. Dr. W. Junk, The Hague.

Wilkinson, R. C., J. W. Hanover, J. W. Wright, and R. H. Flake. 1971. Genetic variation in monoterpene composition of white spruce (*Picea glauca*). For. Sci. 17: 83-90.

OTHER REFERENCES

Austin, M. P. 1968. An ordination study of a chalk grassland community. J. Ecol. 56:739-757.

Burley, J., and P. M. Burrows. 1972. Multivariate analysis of variation in needles among provenances of *Pinus kesiya* Royle ex Gordon. Silvae Genet. 21:69-77.

Cailliez, F., and P. Gueneau. 1972. Analyse en composantes principales des propriétés technologiques de bois Malgaches. Ann. Sci. For. 30:215-266.

Dagnelie, P. 1966. Introduction a l'analyse statistique a plusieurs variables. Biom. Proxim. 7:43-66.

Dagnelie, P. 1971. Some ideas on the use of multivariate statistical methods in ecology. In Patil, G. P., E. C. Pielou, and W. E. Waters (Editors). Statistical Ecol. Vol. 3: Many Species Populations, Ecosystems, and Systems Analysis. Penn. State Univ. Press, University Park, Pa. p. 167-180.

Debazac, E., and R. Tomassone. 1965. Contribution à une Etude comparee des Pins Méditerranéens de la Section Halepensis. Ann. Sci. For. 22:215-250.

Dempster, A. P. 1971. An overview of multivariate data analysis. J. Multivariate Anal. 1:316-346.

Farmer, S. A. 1971. An investigation into the results of principal component analysis of data derived from random numbers. Statistician 20:63-72.

Gnanadesikan, R., and M. Wilk. 1968. Data analytic methods in multivariate statistical analysis. Int. Symp. Multivariate Anal. Dayton, Ohio. p. 593-638.

Gnanadesikan, R., and M. Wilk. 1969. Data analysis methods in multivariate statistical analysis. J. Multivariate Anal. 2:593-638.

Holland, D. A. 1968. The component analysis approach to the interpretation of plant analysis data from groundnuts and sugar cane. Exp. Agric. 4:179-185.

Holland, D. A. 1969. Component analysis: An approach to the interpretation of soil data. J. Sci. Food Agric. 20:26-31.

Kershaw, K. A., and R. K. Sheperd. 1972. Computer display graphics for principal component analysis and vegetation ordination studies. Can. J. Bot. 50:2239-2250.

Moore, C. S. 1965. Interrelations of growth and cropping in apple trees studied by the method of component analysis. J. Hort. Sci. 40:133-149.

Rao, C. R. 1961. Some observations on multivariate statistical methods in anthropological research. Bull. Int. Stat. Inst. 38:99-109.

Rao, C. R. 1964. The use and interpretation of principal component analysis in applied research. Sankhya 26:329-357.

Rao, C. R. 1972. Recent trends of research work in multivariate analysis. Biometrics 28:3-22.

☆U.S. GOVERNMENT PRINTING OFFICE: 1975—668-529/83

Isebrands, J. G., and Thomas R. Crow.
1975. Introduction to uses and interpretation of principal component analysis in forest biology. USDA For. Serv. Gen. Tech. Rep. NC-17, 19 p., illus. North Cent. For. Exp. Stn., St. Paul, Minn.

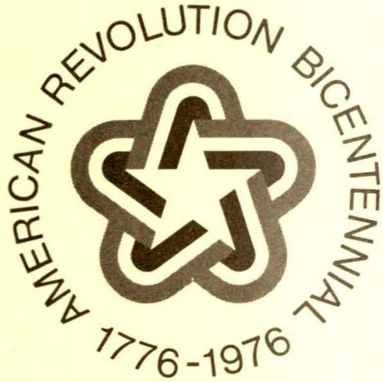
The application of principal component analysis for interpretation of multivariate data sets is reviewed with emphasis on (1) reduction of the number of variables, (2) ordination of variables, and (3) applications in conjunction with multiple regression.

OXFORD: 0--015.5. KEY WORDS: multivariate analysis, eigenvector, principal component, ordination, orthogonal regression.

Isebrands, J. G., and Thomas R. Crow.
1975. Introduction to uses and interpretation of principal component analysis in forest biology. USDA For. Serv. Gen. Tech. Rep. NC-17, 19 p., illus. North Cent. For. Exp. Stn., St. Paul, Minn.

The application of principal component analysis for interpretation of multivariate data sets is reviewed with emphasis on (1) reduction of the number of variables, (2) ordination of variables, and (3) applications in conjunction with multiple regression.

OXFORD: 0--015.5. KEY WORDS: multivariate analysis, eigenvector, principal component, ordination, orthogonal regression.


AMERICAN REVOLUTION BICENTENNIAL
1776-1976